DOCUMENT RESUME

ED 190 089 IR 008 570

AUTHOR Levine, Gwen R.
TITLE The Synergism of User Needs, System Requirements, and Technological Limitations.
PUB DATE Mar 80
NOTE 19p.: Paper presented at the National Online Information Meeting (New York, NY, March 25-27, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Science: *Data Bases: *Data Processing: Information Needs: *Information Retrieval: Information Systems: *Online Systems: Systems Development: *Technological Advancement
IDENTIFIERS *Natural Language: Numeric Data Bases

ABSTRACT

User perception of data importance and the economic feasibility of hardware and storage devices will determine the future direction of online systems. A retrieval system's functions are ultimately dictated by database design. Early online sytems were designed for bibliographic information limited to citations only using sequential files. When indexes and abstracts became desirable for users, their controlled vocabulary requirements could be economically met by inverted files and direct storage devices. The actual users, however, were information specialists familiar with the specialized vocabularies, rather than subject specialists. Now full text information, numeric information, and manipulative systems are being perceived by a class or users not interested in specialized vocabulary skills. These users are placing demands that will require sophisticated internal technology, improved and expanded storage capability, more sophisticated file structure, and transparent data manipulative devices. The economic potential of these demands will bring about the development of a new generation of multitiered databases that allow users at all skill levels to approach them. The point is that, as the user's need for types of data to be retrieved changes, so must the database technology change. (RAA)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED190089

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE SYNERGISM OF USER NEEDS, SYSTEM REQUIREMENTS AND TECHNOLOGICAL LIMITATIONS

by

Gwen R. Levine

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Gwen R. Levine

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

IR008570

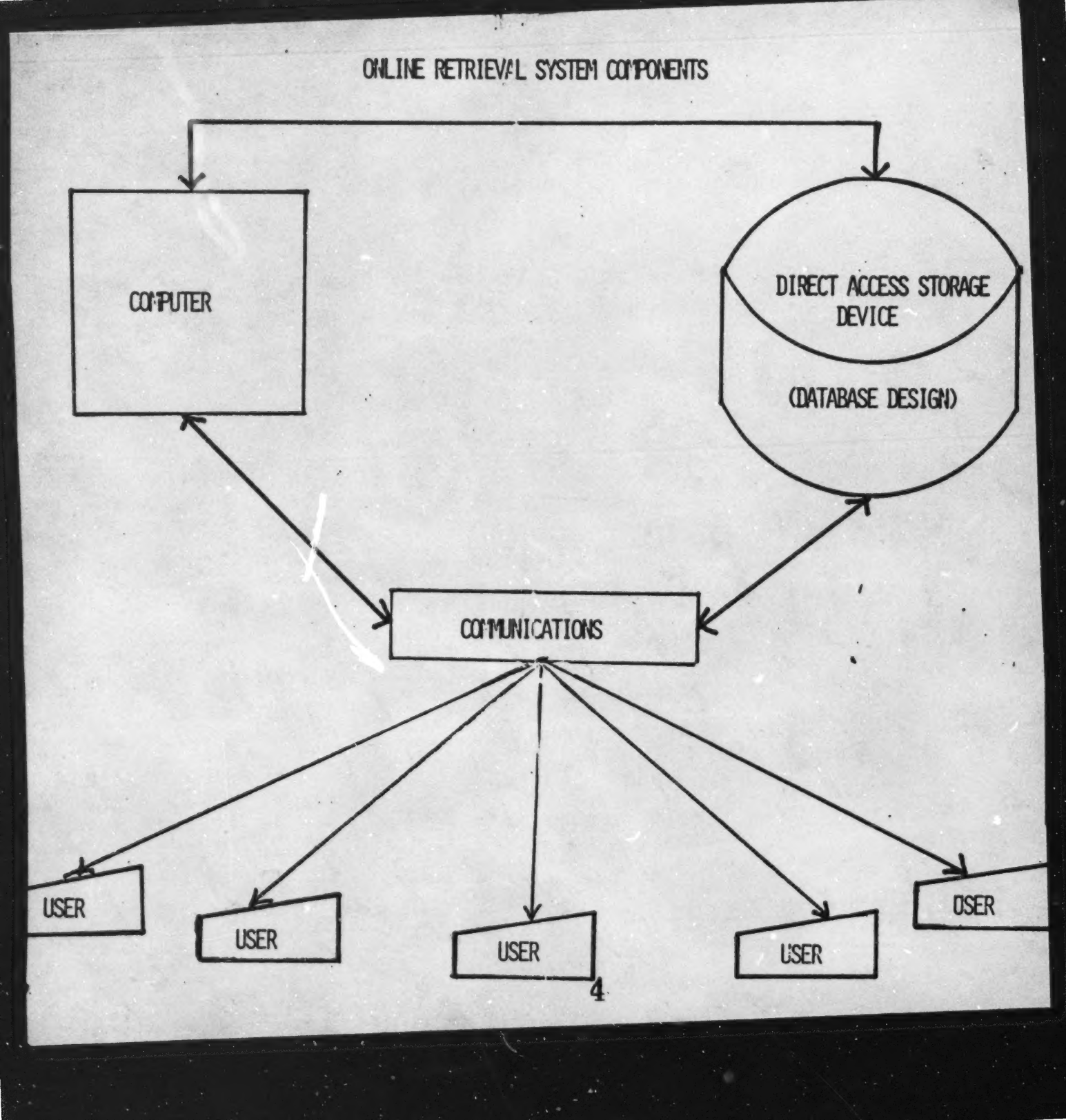
THE SYNERGISM OF USER NEEDS, SYSTEM REQUIREMENTS AND TECHNOLOGICAL LIMITATIONS

BY
GWEN R. LEVINE
SYSTEM DEVELOPMENT CORPORATION
SANTA MONICA, CALIFORNIA

Most discussions of online systems usually revolve around the retrieval program, i.e., how the user interacts with the computer system. However, in the design of an online system the retrieval program per se is not the determining consideration. This is because, ultimately, the retrieval program's functions are dictated by the database design, which has been up to this point, controlled by the hardware devices and storage media available, in other words, by the technology.

Thus, user interfaces become intricately involved with what is perceived to be technically possible and economically feasible. As this perception changes and based on technological changes, systems tend to evolve to fulfill the perception. This is not to say that there is not some ideal way in which a user would prefer to interact with a computer system, it's just that there is a synergistic relationship between user needs, system requirements and technological limitations. As technological problems are solved, a closer approximation of this ideal user interaction can be affected in the system.

F.W. Lancaster wrote in 1973, that one of the major advantages of online systems is that they "can be used directly by the person with the information need without delegating the search to an information specialist". If this is so,why aren't more end users taking advantage of online retrieval fifteen years after its introduction? It seems to me that analyzing the process of database design, rather than the functions of the retrieval program, sheds more light on how we have arrived at the state-of-the-art and what the technical problems are that must be addressed to make any real advances in the so-called end user interface.


ONLINE RETRIEVAL SYSTEM COMPONENTS
COMPUTER
DIRECT ACCESS STORAGE DEVICE
(DATABASE DESIGN)
COMMUNICATIONS
USER
USER
USER
USER
USER

There are three major factors that influence database design: the kinds of informations to be stored and retrieved, the relationship of the data to each other and the physical organization of the information within the computer system. By database I am not referring to any particular collection of data, rather to the database structure,the organization and logical relationship of all of the data elements needed by the computer system in order to perform what is required, some of which are invisible to the system user. This includes the fields to be searched and their linkages to each other; for example, what fields constitute a record, and how are they identified by the system and by the user? How these issues were resolved by the designers of the major online bibliographic systems provides an interesting insight into the synergy of perception of user's needs and the constraints of the possible.

The early online IR systems designed for the Department of Defense, NASA, and, later, the National Library of Medicine were built to handle bibliographic information; not numeric, not textual, not graphic, but bibliographic. Furthermore, the concept of bibliographic data was limited to "citation" information, that is, author, title, source. This conception was a hold over from previous experience with batch and manual retrieval systems.

Because it was not considered possible to operate a manual retrieval system, i.e., printed index or abstracting service, adequately, without a controlled vocabulary, considerable intellectual and financial resources were invested in creating rigorously controlled vocabularies to facilitate retrieval. With the recognition of the information processing power of the computer, batch retrieval systems were developed that basically automated manual retrieval techniques. They were designed around the sequential storage device of magnetic tape, which required the processing of every record in the database to arrive at search results. For this reason, it was still not economically feasible to process anything resembling a full document.

Even when direct access storage devices were able to acheive reasonable economy, the data that was to be retrieved was the by-product of earlier restrictions on what was to be made searchable.

1
AU - FORRESTER DJ
TI - REPRODUCTIVE PERFORMANCE OF A LABORATORY COLONY OF THE DEERMOUSE PEROMYSCUS MANICULATUS GAMBELII.
LA - ENG
MH - ANIMAL EXPERIMENTS/
MH - ANIMALS, LABORATORY/◆PHYSIOLOGY
MH - FEMALE
MH - HOUSING, ANIMAL
MH - LIGHT
MH - LITTER SIZE
MH - MALE
MH - MICE/◆PHYSIOLOGY
MH - PREGNANCY
MH - ◆REPRODUCTION
MH - SEX DETERMINATION
MH - SEX RATIO
MH - UNITED STATES GOVERNMENT SUPPORTED, NON-N.I.H./
CI - 0023-6764 25:446 75
SO - LAB ANIM SCI 25(4):446-9, AUG 75

MEDLINE CITATION

It was the nature of the data to be processed, fielded variable length character data, and the type of storage device suitable for online storage and retrieval-direct access discs- that dictated file organization. It was decided that, for this type of data, an inverted file for the searchable data elements and a sequential file for the printable records was the most efficient database architecture. The problems that might have been addressed by the consideration of other types of data, for example, full text or numerical data,were not part of the original problem definition. The resultant database design had interesting consequences for who would be the real users of the retrieval software systems that exist pretty much intact today.

Davis McCarn and Joe Leiter, in an article in Science in 1973, apparently believed that the real users of the system developed for NLM, would be the medical professional, who could get help if necessary from specially trained search analysts to formulate their search strategies. Although McCarn and Leiter were optimistic that it was possible "to use the system without extensive knowledge" of the NLM medical subject headings, few medical professionals invested the time required to master the vocabulary. Instead the vocabulary 'experts', the medical librarians, became the real users of the ELHILL system. Searching became an "effort to identify how a given document was indexed". The emphasis on controlled vocabulary actually became a barrier to the end users' entry into the system. It was a perfect case of Mooer's law: "an information retrieval system will tend not to be used whenever it is more painful and troublesome for a customer to have information than for him not to have it".

When the ELHILL prototype was implemented as ORBIT in a commercial environment in 1973, in addition to MEDLINE, CHEMCON and ERIC were the first databases to made available. Their record structures were designed pretty much according to the MEDLINE model,although they presented actually quite different problems. CHEMCON had only semi-controlled indexed terms; the controlled indexing would not be available for another five years. ERIC had a loosely

FACTORS IN DATABASE DESIGN

BATCH SYSTEM		ONLINE SYSTEM
SECONDARY PUBLICATIONS	TYPE OF DATA	BIBLIOGRAPHIC CITATION
UNIT RECORDS	RELATIONSHIP BETWEEN DATA ELEMENTS	FIELDED VARIABLE LENGTH RECORDS
SEQUENTIAL FILE (TAPES)	PHYSICAL ORGANIZATION	DIRECT ACCESS DISC - INVERTED FILE (SEARCHING) - SEQUENTIAL FILE (PRINTINT)

controlled vocabulary due to the nature of the discipline, and abstracts as well. As database designers tried to fit more and more databases into the original model of what a record structure should be, it was found that to improve retrieval on those with partially controlled vocabularies, it was necessary to provide additional access to other subject-related fields. Many of these databases had no vocabulary at all, and titles were the only source of subject content. It also became clear that controlled vocabularies would become a hinderance to the broader base of users that were being attracted to online, few of which could be expected to master the vocabularies of the many disciplines represented by the growing list of available databases.

Gradually, what was considered legitimate subject access was expanded, first by making titles and then abstracts searchable in the same way as vocabulary or index terms. Eventually, the Basic Index model evolved, reflecting a relaxation over arbitrarily controlling user searching behavior. The Basic Index model allows searching of subject related vocabulary while requiring no special knowledge of the record structure. This change in approach has required the development of retrieval system features that deal with this change in data element content. Nesting is an example of how a retrieval system allows the construction of quite complex search strategies using single word terms in loose word associations, combined in logical groups, that approximate concepts often as precisely as combining controlled vocabulary terms.

Beyond changing the data element content, the basic record structure has also been modified to handle the growing number of textual files. CIS and ASI were among the first databases which required a new file structure to preserve contextual relationships between main records and subrecords, much as full text systems require tracking of sentence to paragraph to page to chapter to document relationships. This hierarchal file structure is transparent to the user,requiring only the use of a special logical operator, LINK, to invoke the capability. This same file structure has been applied to other databases where contextual relationships exist, for example, the index phrases in CAB and PROMT.

DATABASE
VOCABULARY STRUCTURES

CONTROLLED VOCABULARY ---------- UNCONTROLLED VOCABULARY

(MEDLINE)
APILIT
APIPAT
CHEMDEX
GEOREF
PAPERCHEM
TULSA

COMPENDEX
ERIC
LABORDOC
PSYCHABS

BIOSIS
CAS Files
COLD REGIONS
DERWENT Files
FEDEX
FOREST PRODUCTS
PROMT
SAE
SSIE
USPSD

AGRICOLA
ASI
CIS
LIBCON
LISA
NTIS
TITUS

CAB*
CBPI
CONF. PAPERS
ENERGYLINE
ENVIROLINE
FSTA
INFORM
ISMEC
MANAGEMENT
OCEANIC
POLLUTION
SPORTS

BUSINESS
CDI
CIN
CNI
CRECROD
ELCOM
FEDREG
NDEX
P/E NEWS
SSCI
SAFETY
USCA

CONTROLLED VOCABULARY SEARCH STATEMENT

SS 1 NATIONAL HEALTH PROGRAMS OR COMPREHENSIVE HEALTH PLANNING

SS 2 DELIVERY OF HEALTH CARE OR QUALITY HEALTH CARE

SS 3 1 AND 2

UNCONTROLLED VOCABULARY SEARCH STATEMENT

SS 1 (((HEALTH AND (PROGRAMS OR PLANNING)) AND (NATIONAL OR COMPREHENSIVE)) AND ((DELIVERY OR QUALITY) AND (HEALTH AND CARE)))

-1-
AN - 77-H301-6
TI - Forest Haven.
SO - Apr. 26-27, May 26, 1976.; vii+292 p. foldout.
AV - ^ CIS/MF/5
IS - 7702
CS - 94-2
SI - Y4.D63/1:94-19. Item 1014
MCLC- 77-3991.
RS - 13, 1500, 7600
AB -

Committee Serial No. 94-19. Joint hearings before full committee and Subcom on Education, Labor and Social Services t investigate the quality of staffing and services provided by Forest Haven Institution for the mentally retarded. In 1974, Congress authorized $ 36 million in extra Federal funds to deinstitutionalize residents when appropriate and to upgrade the Forest Haven facility over an 8-year period under the administration of the D.C. Dept of Human Resources (DHR).
Testimony focuses on DHR administration and alleged misallocation of these funds, and on Forest Haven's implementation of D.C. court order requiring the D.C. public schools to provide mentally retarded children with optimum education possible.
Supplementary material (p. 273-292) includes correspondence, submitted responses to subcom questions, and D.C. public schools' FY75 amendment to State Plan (under Part B, Education of the Handicapped Act) providing for full educational opportunity to all handicapped children (p. 276-289).
--*
AN - 77-H301-6.1
ARS - 460
ASO - Apr. 26, 1976. p. 4-19.
WI -
WITNESS: Shackleton, Polly, chairperson, Committee on Human Resources and Aging, DC Council.
AAB -
STATEMENT AND DISCUSSION: Criticism of DHR implementation of 8-year plan objectives and use of available Federal funds; D.C City Council difficulties in exercising oversight of DHR.
--*
AN - 77-H301-6.2
ARS - 524
ASO - Apr. 26, 1976. p. 19-31.
WI -
WITNESSES: Donahue, Michael (Rev.), asst dir, Natl Apostolate for Mental Retardation.; Doe, John, former Forest Haven resident.
AAB -
STATEMENT AND DISCUSSION: Problems in and recommendations for improving Forest Haven deinstitutionalization programs; former resident's description of conditions at Forest Haven.
--*
AN - 77-H301-6.3
ARS - 476
ASO - Apr. 26, 1976. p. 31-43.
WI -
WITNESS: Evans, Betty ACC. BY: Evans, Harold
WITNESSES: Felton, Theressa C., all parents of Forest Haven residents.; Atkins, Bertha N., parent of Forest Haven resident also representing Friends of Forest Haven.
AAB -
STATEMENT AND DISCUSSION: Residents' parents' critical descriptions of Forest Haven conditions.
--*
AN - 77-H301-6.4
ARS - 732
ASO - Apr. 26, 1976. p. 44-63.
WI -
WITNESS: Black, Gorham L., Jr., Regional Dir, Region III, HEW; ACC. BY: Taylor, Robert J., Regional Dir, Long Term Care, Standards Enforcement;Koster, Harvey, Asst Regional Commr for Mgmt, Social and Rehabilitation Service;Marsicano, Rosan, Office of Long Term Care, Standards Enforcement.
AAB -
STATEMENT AND DISCUSSION: Description of HEW medicaid program eligibility standards for intermediate care of the mentally retarded; chronology and findings of investigations resulting in Forest Haven loss of medicaid eligibility (tables, p. 50-51).
--*
AN - 77-H301-6.5
ARS - 388
ASO - Apr. 26, 1976. p. 63-69.

The point is that, as the type of data to be retrieved changes, so must the file structures that support them. This has a direct impact on the user interface, requiring the retrieval program to expand its repertory of functions. Although existing online systems can continue to make modifications to handle textual data fairly well, there will have to be major database design adjustments to allow for the handling of numeric and properties databases allowing for file structures that are more appropriate for that kind of data. Also in order to utilize this type of database effectively, new capabilities will be developed including statistical data manipulation, special display functions including graphics, and more versatile output report generation. To make development of these types of files commercially viable it will be necessary to attract users who have the need to use them and on any broad scale that will have to be the end user.

As I have tried to demonstrate, we have reached the stage of evolution in existing online systems that certain technical problems must be overcome before more"friendly" end user interfaces can be developed. The broadening of vocabulary access, along with tools to deal with this kind of approach, is a step in this direction, as is the ability to handle full text files. The lesson to be learned from the trend away from controlled to uncontrolled vocabulary searching, is that the ideal online system to be used by the subject specialist who hasn't the time to invest in learning either a special vocabulary or system commands is one that approximates natural language interaction. There are at least two major hurdles that must be cleared before systems can approach this ideal.

A natural language searching system requires access to the document itself, not just an abstract or indexed citation since the vocabulary of the original text coincides with that of the end user. We have seen that controlled vocabularies are the domain of the information specialist whereas natural language systems favor the subject specialists. The most obvious problem here to a system that provides access to over 20,000,000 citations is storage on such a massive scale. It may be that some kind of networking scheme of storage at remote sites such as at the

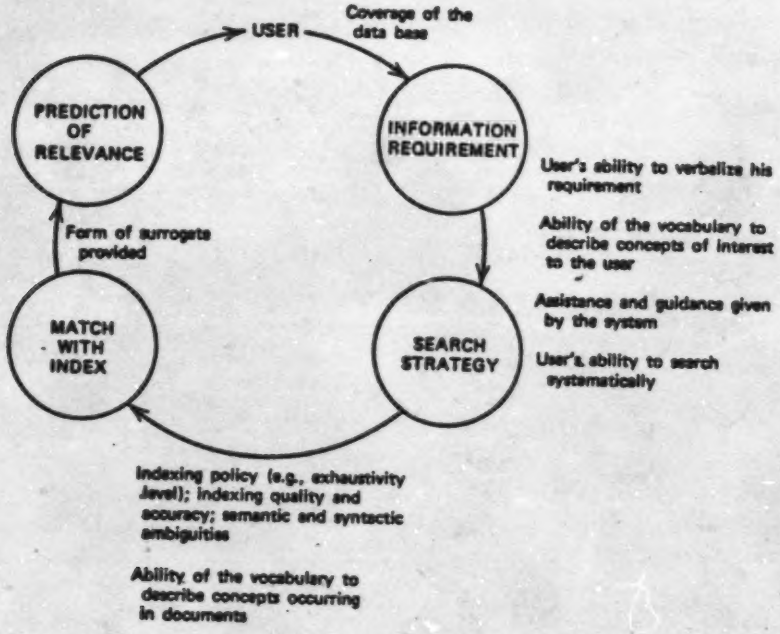
publishers and retrieval by a central computer might be a possible future scenario.

Another major barrier to the development of natural language systems is the structure of language itself which creates giant new problems for database design. The invisible vocabulary aids that would be necessary in such a system would be very costly to store and manipulate. Consider, for example, how would the data elements which represent the entire vocabulary of a language be related? What kind of logical file structures can assure that the document is really about a topic or simply contains a tangential reference to it? The consideration of the problems represented by this type of data must await the emergence of economically feasible storage devices that will in turn determine to a large extent what the user interface will be.

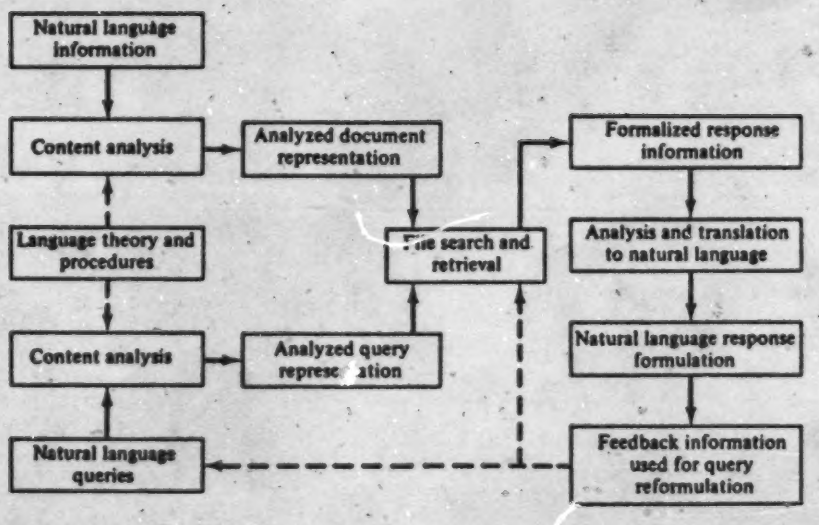
More research directed toward the development of systems that process language not data is the crux of the matter- a problem many levels of magnitude more difficult than existing systems. Note the relative simplicity of inverted index searching in the viewgraph as opposed to the complex language analysis modules required by the hypothetical natural language system.

The next generation of systems will be built on multi-tiered databases that allow users of all skill levels to approach them on their own terns, structured or unstructure, depending on their individual needs. These structures will become increasingly transparent to the user, and the systems will become 'smarter' translating user requests into 'best fit' strategies depending on the data content of a particular database or set of databases. They will have many more built-in default conditions such as compensating for common human errors, e.g., spelling errors. Certain situations will trigger menus that guide the user through possible options. Thesaurus aids such as automatic term explosions or synonymn lists built and used by the user, and multilingual thesauri translation will be components of the database designs of the future.

The process of getting there will be the same as has brought us this far. Depending on what data are perceived to be important for


USER
Coverage of the data base
PREDICTION OF RELEVANCE
INFORMATION REQUIREMENT
User's ability to verbalize his requirement
Ability of the vocabulary to describe concepts of interest to the user
Assistance and guidance given by the system
User's ability to search systematically
Form of surrogate provided
MATCH WITH INDEX
SEARCH STRATEGY
Indexing policy (e.g., exhaustivity level); indexing quality and accuracy; semantic and syntactic ambiguities
Ability of the vocabulary to describe concepts occurring in documents

RETRIEVAL USING INVERTED INDEXES



RETRIEVAL USING HYPOTHETICAL LANGUAGE ANALYSIS SYSTEM

online access, and what hardware and storage devices are considered to be economically feasible, systems will tend to be created to fulfill these perceptions.

REFERENCES

F.W. Lancaster, Information Retrieval On-Line, Melville Publishing Company, Los Angeles, California, 1973, ISBN 0-471-51235-4.

Davis B. McCarn and Joseph Leiter, "On-Line Services in Medicine and Beyond", Science, July 27, 1973, Vol. 181, p. 318-324.

C.N. Mooers, "Mooers' Law, or Why Some Retrieval Systems Are Used and Others Not", American Documentation, 1960, Vol. 11, No. 3, ii.

Gerald Salton, Dynamic Information and Library Processing, Prentice-Hall, Englewood Cliffs, New Jersey, 1975, ISBN 0-13-221325-7.

Donald D. Spencer, Introduction to Information Processing, 2nd ed., Charles E. Merril, Columbus, Ohio, 1977.